Can a Robot Have Free Will?

Farnsworth, K. (2017). Can a Robot Have Free Will? *Entropy*, *19*(5), [237]. https://doi.org/10.3390/e19050237

Published in:
Entropy

Document Version:
Publisher's PDF, also known as Version of record

entropy

MDPI

Article

Can a Robot Have Free Will?

Keith Douglas Farnsworth

School of Biological Sciences, Queen's University Belfast, Belfast BT97BL, UK; k.farnsworth@qub.ac.uk; Tel.: +44-2890-97-2352

Academic Editor: Mikhail Prokopenko
Received: 27 February 2017; Accepted: 15 May 2017; Published: 20 May 2017

Abstract: Using insights from cybernetics and an information-based understanding of biological systems, a precise, scientifically inspired, definition of free-will is offered and the essential requirements for an agent to possess it in principle are set out. These are: (a) there must be a self to self-determine; (b) there must be a non-zero probability of more than one option being enacted; (c) there must be an internal means of choosing among options (which is not merely random, since randomness is not a choice). For (a) to be fulfilled, the agent of self-determination must be organisationally closed (a "Kantian whole"). For (c) to be fulfilled: (d) options must be generated from an internal model of the self which can calculate future states contingent on possible responses; (e) choosing among these options requires their evaluation using an internally generated goal defined on an objective function representing the overall "master function" of the agent and (f) for "deep free-will", at least two nested levels of choice and goal (d–e) must be enacted by the agent. The agent must also be able to enact its choice in physical reality. The only systems known to meet all these criteria are living organisms, not just humans, but a wide range of organisms. The main impediment to free-will in present-day artificial robots, is their lack of being a Kantian whole. Consciousness does not seem to be a requirement and the minimum complexity for a free-will system may be quite low and include relatively simple life-forms that are at least able to learn.

Keywords: self-organization; downward causation; autocatalytic set; goal-oriented behaviour; autopoiesis; biological computing

1. Introduction

Why do things do what they do? We have a hierarchy of explanations that roughly reflects a gradient in complexity, matched by the epistemic hierarchy which starts with the physics of Hamiltonian mechanics (and Shrödinger's equation), extends through statistical mechanics and complex systems theory, but then declines in power as we try to account for the behaviour of living systems and finally of the human condition, for which we have no satisfactory scientific explanation. One of the most persistent open questions, at the far end of the complexity gradient, is whether we as humans have free will. Here, I attempt to address this question with respect to a broader category of active agent (sensu Sharov [1]): anything that can make decisions and act in the physical world. The premise of this paper is that an understanding of the interaction and dynamics among patterns, of the distribution of matter and energy in space and time, may bring such high-level phenomena into resolution. That means a focus on information and its interactions using cybernetics and computation theory, but it also requires a broad concept of information addressing the relationship among patterns (as data) in general, rather than just the statistics of data transmission. If the emergence of material reality (as we experience it) is the assembly of stable configurations (of matter), undergoing transformations and combining as stable composites (e.g., molecules forming materials, forming structures [2]), then a deep understanding of it requires a mathematically precise account of the physics of patterns that are simultaneously the product of material structure and the cause of it [3]. For this

investigation, we must focus on the kind of "information" that is embodied by, and processed by natural systems. This concept of intrinsic structural "information" (surely it ought to have a word of its own), is not necessarily ontological (as in the theories of Weizsäcker [4] and Stonier [5]), but refers to at least observable patterns having observable effects [6], is objectively quantifiable [7] and useful in understanding biological processes in terms of cybernetic systems [1], and functions [8,9]. To avoid ambiguity (and conflict) I will refer to this kind of "information" as pattern and the "information" which reduces "uncertainty" in a receiver as Shannon information.

My aim is to discover whether free will can be understood in terms that make no reference to specifically human qualities and whether free will could be attributed to a broader class of active agent (sensu Sharov [1]). I aim to do this by identifying the criteria that must be met for free will, and from that, to identify the kinds of active agent which might meet those criteria.

I will start by defining what I mean by "free-will" and then give a very brief overview of the philosophical debate about free-will, identifying some of the problems. Then I will introduce ways of thinking about these problems based on cybernetic/computation theory and use these to identify the necessary and sufficient conditions for free-will according to my definition. The class of systems for which these conditions are fulfilled will then be identified and with this, the minimum complexity compatible with autonomous action will be implied. The term "robot" is specifically used here to mean any cybernetic system coupled to a physical system that allows it to independently act in the physical world; this includes living systems (see e.g., the requirements specified for a molecular robot in Hagiya et al. [10]) and artificial systems, including the subject of cognitive robotics which concerns embodied artificial intelligence [11].

1.1. A Definition of Free-Will

There is no generally agreed definition for free will. To find the conditions allowing active agents (sensu Sharov [1]) to have free-will, a working definition is first provided. It is defined here as the condition in which all of the following are jointly true:

- FW1: there exists a definite entity to which free-will may (or may not) be attributed;
- FW2: there are viable alternative actions for the entity to select from;
- FW3: it is not constrained in the exercising of two or more of the alternatives;
- FW4: its "will" is generated by non-random process internal to it;
- FW5: in similar circumstances, it may act otherwise according to a different internally generated "will'.

In this definition the term "will" means an intentional plan which is jointly determined by a "goal" and information about the state (including its history) of the entity (internal) and (usually, but not necessarily) its environment. The term "goal" here means a definite objective that is set and maintained internally. The terms used here will be explained and justified in what follows.

This list of criteria is chosen to address the main features that most philosophers have thought important (though they do not necessarily agree with one another about what is important and some philosophers would leave out some items of the list). FW2 and FW5 are intended to examine the effect of determinism and FW4 represents the "source arguments" for and against free will, whist FW3 ensures freedom in the most obvious (superficial) sense. Only one of the list (FW1) is not usually included in any philosophical discussion of free will, perhaps because it is usually considered to be self evident, but it will play an important role here.

1.2. The Philosophical Background

In philosophy, freedom "to act as one wills" is often referred to as "superficial freedom" (or first order volition, [12]) and the freedom "to determine what one wills" as "deep freedom" (or second order volition, [12]). Kane [13], McKenna [14] and Coeckelbergh [15] provide a broad introduction to the subject, Westen [16] gives a deeper criticism of it. As an example, one may be free to drink a bottle

of whisky in one sitting and as an alcoholic, or in an irresponsible mood, or emotional turmoil one may will it, but knowing the consequences, one may master the desire and will otherwise: rejecting the opportunity of this poisonous pleasure. If the alcoholism had taken over, or one was under duress, the "freedom to choose otherwise" might be denied and deep freedom would be lost with it, although the superficial freedom to drink would remain. The classical philosophical argument on free-will consists of a) whether it is compatible or not with determinism and b) whether at least some agents (usually people) are at least in part the ultimate cause of their actions.

Determinism is the idea that there is, at any instant, exactly one physically possible future [16,17], summarised in the slogan "same past: same future" (see List [18] for a more rigorous analysis). Cybernetics captures determinism in the definition of a Determinate Machine (DM) as a series of closed, single valued transformations (for example describing a Finite State Automaton (FSA)). Superficial freedom is often seen as the absence of constraint, leading to the (relatively trivial) conclusion that it is compatible with determinism. However, deep freedom needs more than an absence of constraints, it requires alternative paths into the future to provide the "freedom to do otherwise" [17]. The cybernetic model of a system with this capacity is of course the non-determinate machine (NDM). However, the NDM is usually conceived as a probabilistic process in which a set of possible states $\mathrm{S} : \{s_1, ..., s_n\}$ of the system, given the present conditions C (in general including the previous history), may occur at random, with probability set $\mathrm{P} : \{p_1, ..., p_n\}$ where $\sum p_i = 1$ and each p_i is the probability of each possible state s_i. Most philosophers agree that randomness is not compatible with self-determination, indeed it seems to be the opposite, so they reject random spontaneity as a means of achieving deep freedom (and so do I). They reason that if an agent's action were ultimately caused by e.g., a quantum fluctuation or thermal noise, then we could not reasonably hold the agent responsible for it. This indicates that philosophers supporting the existence of deep free-will are searching for a non-random ultimate cause of actions within the agent of those actions.

Unfortunately for them, a paradox arises: since any agent is the product of its composition and of its previous experiences (which may play a role in its formation) and these are beyond its control. If it did not make itself and select its own experiences, then its behaviour must be determined by things other than itself. An agent is not free in the deep sense unless it has control over all the events that led to it's choice of action. Recognising that all events in the universe belong to a chain of cause and effect that extends back before the existence of the agent, some philosophers conclude that either (a) this deep freedom cannot exist and is considered an illusion (e.g., Van Inwagen [17], reiterated in [19]); or (b) the agent is indeterminate so that we get "same past: different futures". If they also rule out randomness, then (b) suggests that an agent which could act in more than one possible way from exactly the same state and history (i.e., it is indeterminate) must act without cause. They conclude that this is self-contradictory, hence deep freedom cannot exist. This line of thinking is closely related to Strawson's [20] "Basic Argument" against free-will, which starts from the premise that for an agent to have free-will it must be the cause of itself and shows, via infinite regress, that this is not possible. It is an axiom of these positions that the motivation for an agent's action can only be either (a) random; (b) exogenous or (c) self-generated, with only (c) being compatible with free-will.

2. Systems with Identity: The Closure Condition

The first criterion (FW1) implies that we need to determine what parts of a system must be included in identifying an agent (i.e., the extent of its identity) before being able to determine if it has free-will or not. Free-will requires a definite boundary between the internal and external, not necessarily a physical boundary (as supplied by e.g., a casing or skin), but more profoundly one of organisation and control. For example a computer controlled robot must have all the necessary provisions for physical independence (as in the extraterrestrial exploration robots), but this still leaves it organisationally linked to humanity because its existence is entirely dependent on our gathering and processing the materials for its "body" and assembling these, implicitly embodying it with functional information [1] and programming its control computer (including with the goal for operation). For these reasons, such

robots remain extensions of ourselves: tools just as sophisticated hammers would be. In general, for free-will, the control information of an agent must be independent of anything beyond a cybernetically meaningful boundary. Put the other way round, for the identification of free-will, we must first identify the boundary of the agent, which is defined by independence of control. The existence of such a cybernetic boundary enclosing the agent is here termed the "closure condition". Given this, the Mars Rover coupled with its human design team seems to meet the closure condition, but the Mars Rover alone does not.

This idea of a boundary surrounding a system, such that whatever is within the boundary has the property of organisational independence from what lies without, was encapsulated by the concept of the "Kantian whole" by Kauffman [21] and can be formally described in cybernetic terms as organisational closure. This idea is used in a particular definition of autonomy which I shall next argue to be a pre-requisite for free will.

2.1. Systems with Causal Autonomy

Froese et al. [22] distinguishes behavioural (based on external behaviour) from constitutive (based on internal organisation) autonomy. They state that for the former, the identity of the system may be imposed by an external observer (it could even be no more than an thermostatic system) and that it is sufficient for the system to demonstrate a capacity for stable and/or flexible interaction with its environment, "without human intervention, supervision, or instruction". This, they argued, left behavioural autonomy so "ambiguous and inclusive . . . it threatens to make the concept of autonomy meaningless". Behavioural definitions of autonomy are inadequate in relation to questions of free will because they do not address the source of will (more precisely, the origin of the goal/s) which motivate the observed actions. To attribute free will to an agent, we need to identify the source of will as a part of the agent (under FW4) and this requires us to consider the composition of the agent. To exclude external (generalised from "human") "intervention, supervision, or instruction", we must have an agent that is separated from external causation of its actions. According to most authors seeking to explain the apparent independence of action found among living systems, this requirement leads directly to constitutive autonomy, e.g., "every autonomous system is organizationally closed" (Varela [23], p. 58). This idea has a relatively long history in a multi-disciplinary literature (Froese et al. [22], Zeleny [24], Rosen [25], Vernon et al. [26], Bich [27] and references therein), but it is not clear if it is restricted to living systems, or may be broader. Therefore, rather than taking this literature as sufficient justification for a constitutive autonomy requirement, let us examine the options for matching with the following tasks:

- to answer Strawson's [20] "Basic Argument" of ultimate responsibility;
- to separate internal from external (i.e., to give formal meaning to internal and external);
- to unambiguously break the physical causal link between the agent and its environment, allowing "leeway" from determinism.

This will start with task (2) because from its conclusion, task (3) may follow, given a specification that relations are strictly causal and if the answer to task (2) does specify constitutive autonomy, then task (1) may be implied from that, though it leaves unproven that a causally autonomous system must be self-made. This gap may not be serious, as later shown.

Therefore, we seek a structure for which "internal" is causally distinct from "external", giving a clear definition to both. For this we need to define an object A, properly composed of parts (e.g., x, y), none of which is a part of any other object B. More formally: $\exists$ A composed of parts $a_i \in$ a : $\{a_1 ... a_N\}$ in which no part of A is also a part of any object B unless B contains or is A: $(\mathrm{B} \supseteq \mathrm{A})$. Assume the mereology in which the reflexivity and transitivity principles are true and also the antisymmetry postulate is true. That is: two distinct things cannot be part of each other. This is expressed by the axioms [28]:

- $x \mathrm{P} x$

- $(x\mathrm{P}y \wedge y\mathrm{P}z) \rightarrow x\mathrm{P}z$
- $(x\mathrm{P}y \wedge y\mathrm{P}x) \rightarrow x = y$,

in which P is the "part of" relation. For these, the auxiliary relations are defined:

- Overlap: $x \circ y := \exists z(z\mathrm{P}x \wedge z\mathrm{P}y)$,
- Exterior: $x \perp y := x\neg \circ y$.

Therefore, we can define an isolated object y : no part of y can overlap with anything other than y. Hence, Isolation: $\forall z \;\; z\neg \circ y$, hence $\forall z \;\; z \perp y$ (this definition is not a part of standard mereology). The mereological sum (noting that alternative definitions exist) is defined as: an object y is a mereological sum of all elements of a set X iff every element of X is a part of y and every part of y overlaps some element $x \in \mathrm{X}$. (Effingham [29], p. 153) puts it this way: "the xs compose y by definition if (i) each x is a part of y; (ii) no two of the xs overlap; (iii) every part of y overlaps at least one of the xs". This definition of sum allows overlap with objects that are not parts of the mereological sum, so y is not necessarily isolated, therefore also let $|y(\mathrm{X})|$ denote that y is an object that is both isolated and composed of a set X (of xs).

So far, partness (the $x\mathrm{P}y$ relation) has not been defined. The definition of partness depends on what condition must be met for objects to be associated as parts. This is the Van Inwagen [30] special composition question: "what constitutes being a part; what connection or relationship qualifies as 'partness'?" Specifically, we need a criterion for "restricted composition" that is relevant to the question of separating internal from external in terms of causation. For this, define causation as a binary relation: $x\mathrm{C}y$, specifying that the state of object y is strictly determined by the state of object x. With this, we can define a transitive closure for causation (transitive closure is the minimal transitive binary relation R on a set X).

First, note that any relation R on a set X is transitive iff $\forall a, b, c \in \mathrm{X}$, whenever $a\mathrm{R}b$ and $b\mathrm{R}c$ then $a\mathrm{R}c$. The conditions for a transitive closure can then be written as follows [31] (using the notation R^+ to represent a transitive relation): (i) if $\{a, b\} \in \mathrm{R} \;\rightarrow \{a, b\} \in \mathrm{R}^+$ (ii) if $\{a, b\} \in \mathrm{R}^+ \wedge \{b, c\} \in \mathrm{R} \rightarrow \{a, c\} \in \mathrm{R}^+$ (iii) nothing is in R^+ unless by (i) and (ii).

Next, if R is specified as the C relation (from above), then (iii) specifies all the causal relations among members of a set C^+, so that no relation $\{x, a\} \notin \mathrm{C}^+$ can be causal. This has the effect of causally isolating the elements of C^+ as well as ensuring that they are causally related to one another.

Finally, let Y be the set of elements that are included in a transitive causal closure C^+, where $x\mathrm{C}y$ is the condition for association as parts: x is a part of y. Under these restrictions, mereologically, the elements of Y are the sum of an object y and $|y(\mathrm{Y})|$. This means that the transitive closure for causation meets the requirement for formally separating internal from external (task 2). Since the relations in the closure are defined to be causal, this achieves task 3 as well. An object (agent) with the property of transitive causal closure among its parts is a causally autonomous system. All that is missing is a way to ensure that the agent is a cause of itself from the beginning of its existence, which requires it to be self-constructing. That will next be addressed with a more concrete example.

2.2. The Kantian Whole as a Material System

A system composed of parts, each of whose existence depends on that of the whole system is here termed a "Kantian whole", the archetypal example being a bacterial cell [32]. The origin of this terminology lies in Immanuel Kant's definition of an organised whole [33]. To make the closure condition concrete and include an answer to Strawson's [20] "Basic Argument" it will now be narrowed to a requirement for self-construction, since this implies the embodiment of self with the pattern-information that will then produce the agents behaviour (i.e., we require strictly constitutive autonomy as defined by Froese et al. [22]). In other words, we are to consider a cybernetic system that, by constructing itself materially, determines its transition rules, by and for itself (material self-construction may not be essential to ensuring self-determination, but assuming that the cybernetic relations embodied in it are essential, we may proceed without loss of generality). An autocatalytic

chemical reaction network with organisational closure (and this is also what Kauffman [21] considered a Kantian whole) is an anabolic system able to construct itself [34].

Hordijk and Steel [34] and Hordijk et al. [35] define their chemical reaction system by a tuple $Q = \{X, \mathcal{R}, C\}$, (their symbols) in which X is a set of molecular types, $\mathcal{R}$ a set of reactions and C a set of catalytic relations specifying which molecular types catalyse each member of $\mathcal{R}$. The system is also provided with a set of resource molecules $F \subseteq X$, freely available in the environment, to serve as raw materials for anabolism (noting that whilst we are defining an organisational closure, we may (and indeed must) permit the system to be materially and thermodynamically open). The autocatalytic set is that subset of reactions $\mathcal{R}' \subseteq \mathcal{R}$, strictly involving the subset $X' \subseteq X$, which is:

- reflexively autocatalytic: every reaction $r \in \mathcal{R}'$ is catalysed by at least one of molecular type $x' \in X'$ and
- composed of F by $\mathcal{R}'$: all members of X' are created by the actions of $\mathcal{R}'$ on $F \cup X'$.

This definition of an autocatalytic set is an application of the broader mathematical concept of closure and more specifically of transitive closure of a set, since when the autocatalytic set is represented as a network (of reactions), this network has the properties of transitive closure. The concept of autocatalytic set has been implemented in experiments for exploring aspects of the origin of life (e.g., the GARD system simulating "lipid world" [36]). Clearly with the two conditions for an autocatalytic set met, everything in the system is made by the system, but there is a more important consequence. The system is made from the parts (only) and can only exist if they do. Organisational closure of this kind has been identified as a general property of individual organisms [37], many biochemical sub-systems of life [21] and embryonic development [38].

As it is defined above, living systems fulfil the closure condition, but can we conceive of a non-living system also reaching this milestone? Von Neumann's [39] self-replicating automata show that some purely informational (algorithm) systems have the capacity to reproduce within their non-material domain, but they cannot yet assemble the material parts necessary, nor can they build themselves from basic algorithmic components (they rely on a human programmer to make the first copy). What is needed for the physical implementation of a Kantian whole is the ability to "boot-strap" from the assembly of simple physical components to reach the point of autonomous replication (i.e., the system must be autopoietic [37,40]. This is necessary to answer Strawson's [20] "Basic Argument": that for deep free-will an entity must be responsible for shaping its own form and it provides a motivation for rejecting dualism (the idea that the "mind" is not created from the material universe).

2.3. Emergence and Downward Causation

Considering the forgoing, we might ask what is responsible for making an autopoietic system (e.g., an organism); is it the components themselves, or is it the organisational system. We might further ask: in either case, what really is the "system". Cybernetics provides an answer to the second question, in that the system is the organisational pattern-information embodied in a particular configuration of interactions among the component parts. Because it is abstract of its material embodiment, it is "multiply realisable", i.e., composed of members of functionally equivalent parts (see Auletta et al. [41]; and Jaeger and Calkins [42] for biological examples). It is not the identity of the components that matters, rather it is the functions they perform (e.g., a digital computer may be embodied by semiconductor junctions, or water pipes and mechanical valves, without changing its identity). Crucially, "function" is defined by a relationship between a component and the system of which it is a part. According to Cummins [43], "function" is an objective account of the contribution made by a system's component to the "capacity" of the whole system. At least one process performed by the component/s is necessary for a process performed by the whole system. This implies that the function of a component is predicated on the function of the whole. This definition was recently modified to more precisely specify the meaning of "capacity" and of whole system, thus: "A function

is a process enacted by a system A at organisational level L which influences one or more processes of a system B at level $L+1$, of which A is a component part" [44].

In this context, organisational level means a structure of organisation that is categorically different from those above and below in the hierarchy because it embodies novel functional information (levels may be ontological or merely epistemic in meaning: that is an open debate in philosophy). The self-organisation of modular hierarchy has been described as a form of symmetry-breaking phase transition [45], so the categories either side are quantitatively and qualitatively different. Organisational levels were defined precisely in terms of meshing between macro and micro dynamics (from partitioning the state-space of a dynamic system) by Butterfield [46] and also using category theory to specify supervenience relations and multiple-realisability among levels by List [47]. Neither definition, though, deals specifically with the phenomenon of new pattern-information "emerging" from the organisation of level L components at level $L+1$, which is responsible for the emergence of new phenomena.

Ellis [48] shows that a multiply realisable network of functions, self-organised into a functional whole, emerges to (apparently) exercise "downward causation" upon its component parts [48,49]. The organisational structure is selecting components from which to construct itself, even though it is materially composed of only the selected components. Since it is purely cybernetic (informational) in nature, the downward control is by pattern-information [42,48] which transcends the components from which it is composed. The pattern-information arises from, and is embodied by, the interactions among the components, and for these reasons it was termed a "transcendent complex" by Farnsworth et al. [50]. Examples are to be found in embryonic development, where a growing cluster of cells self-organises using environmental signals created by the cells taking part [51] and the collective decision making of self-organising swarms (e.g., honey bees in which the hive acts as a unity [52]).

There is something significant here for those who conflate determinism with causation. All causal paths traced back would be expected to lead to the early universe. Despite the appearance of near maximum entropy from the uniformity of background microwave radiation, there is broad agreement that the entropy of the early universe was low and its embodied pattern-information (complexity) could not account for the present complexity, including living systems [53]. Novel pattern-information has been introduced by selection processes, especially in living systems, for which Adami et al. [54] draw the analogy with Maxwell's demon. Selection is equivalent to pattern matching, i.e., correlation, and is accompanied by an increase of information. Since its beginning, the entropy of the universe has been increasing [53] and some of this has been used as a raw material for transformation into pattern-information. This is achieved by creating the "order" of spatial correlation through physical self-assembly (atoms into molecules into molecular networks into living systems). This self-assembly embodies new information in the pattern of a higher level structure through the mutual provision of context among the component parts [3]. The process of self-assembly is autonomous and follows a boot-strap dynamic, so it provides a basis for answering Strawson's [20] "Basic Argument" in which the putative agent of free-will is an informational (pattern) structure of self-assembly.

2.4. Purpose and Will

Much of the literature on downward causation uses the idea of "purpose", though many are uncomfortable with its teleological implication. The aim of this section is to form a non-teleological account of purpose and its connection with will in non-human agents.

Cause creates correlation (usually, but not necessarily, in a time-series): the pattern of any action having a cause is correlated with its cause. An action without cause is uncorrelated with anything in the universe and accordingly considered random. If an action is fully constrained, then its cause is the constraint. Thus, freedom from at least one constraint allows the cause to be one of either: random, or exogenous control or agent control (in which "control" means non-random cause). By definition the cause is only taken to be the agent's will if it originated in agent control. Correlation alone, between

some outcome variable x and some attribute a of the agent, is not sufficient to establish will: (a) because correlation has no direction (but metrics such as "integrated information" [55] can resolve direction) and (b) because a may itself be random in origin and thereby not of the agent's making. Marshall et al. [56] showed that cause can be established at the "macro" level of agent (as opposed to the "micro" level of its components) using an elaboration of integrated information, so the pattern in x can be attributed to agent-cause. Because the agent-based cause could be random ((b) above), we must form and test a hypothesis about the effect of x on the agent before we can attribute the cause to the alternative of agent-will. The hypothesis is that the effect of a on x is to increase the overall functioning F of the agent. If this were true, then to act wilfully is to reduce the entropy of x, by increasing the probability of an outcome x' where $x' \Rightarrow F' > \bar{F}$ (and $\bar{F}$ is the average F). That means that the mutual information between a wilful action $a(t)$ at time t and the resulting function $F(t+\tau)$, $\tau \geq 0$ is greater than zero. This mutual information between action and future functioning is taken to imply a "purpose" for the action, so purpose is identified by the observations that:

$$H(a) + H(F) > H(a|F) \text{ and } F|a > F|r, \tag{1}$$

where r is a random (comparator) variable. This is clearly an observational definition and is in some way analogous to the Turing test, but it is a test for purpose rather than "intelligence". It represents our intuition that if a behaviour repeatedly produces an objectively beneficial outcome for its actor, then it is probably deliberate (repeatedly harmful behaviour is also possibly deliberate, but all such actions are regarded as pathological and thereby a subject beyond the present scope).

To recap, for attributing the action to the agent's will, we must at least identify a purpose so that Equation (1) is true. The purpose is a pattern embodied in the agent, which acts as a template for actions of the agent that cause a change in future states (of the agent, its environment, or both). We may call this pattern a "plan" to attain an objective that has been previously set, where the objective is some future state to which the plan directs action. Specifically, let the objective be a state X (of the system or the world, etc.), which can be arrived at through a process P from the current state Y, then the purpose is a "plan" to transform $\mathrm{Y} \rightarrow \mathrm{X}$ by the effect of at least one P and at least one function F is necessary for the process P to complete.

The homoeostatic response to a perturbation, for example, has maintenance at the set-point as its purpose. Y is the perturbed state, F is some function of the internal system having the effect of causing a process P, i.e., some transition $\mathrm{Y} \rightarrow \mathrm{X}$. In general there is more than one P and more than one F for achieving each. This results in a choice of which to use: it is a choice for the agent described by the system. To make a choice requires a criterion for choosing (else the outcome is random and therefore not a choice). The criterion for choosing is a "goal" G, consisting of one or more rules, which identify a location in a function describing the outcome (which we may call the objective function). In general, this location could be any and it is essential for freedom of will that it be determined by the agent of action alone. However, in practice it is most likely to be an optimisation point (in living systems, this is implied by Darwinian evolution and in designed systems, it is the basis of rational design). Therefore, narrowing the scope, but with justification, let us take the criterion for choosing to be a goal G, consisting of one or more optimisation rules. For example, of all the possible systems performing homoeostasis, the purposeful one is defined as enacting P′ such that $\mathrm{Y} \rightarrow \mathrm{X}$, with $\mathrm{P}' \in \mathrm{P}|\max(\mathcal{G})$, where $\mathcal{G}$ is an objective function for which the optimisation goal $\mathrm{G}(\mathcal{G})$ is satisfied, contingent upon the options (e.g., P proceeds as quickly as possible, or with minimum energy expenditure, etc.). Accordingly, "will" is defined by a purpose which is a plan to enact a process causing a transition in state, "as well as possible" (according to $\mathrm{G}(\mathcal{G})$), notwithstanding the earlier comment about pathological purposes. A free-will agent has a choice of transition and a goal which identifies the most desirable transition and the best way to enact it, from those available. These two choices can be united (by intersection), without loss of generality, to one choice of best transition.

One of the reasons for objecting to teleological terms such as "purpose", "plan" and "goal" in relation to natural systems has been the belief that a plan implies a "designer", the concept at the

centre of the most famous battles between science and religion. This implication is not necessary and is rejected here (following the argument of Mayr [57]). A plan is merely a pre-set program of steps taking the system from Y to X; it is the concept for which computation theory was developed. It may be designed (the work of an engineer), but also may have evolved by natural selection (which also supplies the goal, in which case it is a teleonomic system (sensu Mayr [57]).

A plan, as an ordered sequence of transformations, is an abstraction of information from the physical system, which for free-will must be embodied within the system. A more subtle implication of "plan" is that as a path leading from Y to X, it is one among several possible paths: different plans may be possible, perhaps leading to different outcomes. There is a fundamental difference between this and the inevitability of a dynamic system which follows the only path it may, other than by the introduction of randomness. The reason is that for a dynamical system all the information defining its trajectory is pre-determined in the initial (including boundary) conditions and the laws of physics. The initial conditions constitute its one and only "plan". If a system embodies pattern-information (by its structure) which constitutes a developed plan, then this pattern-information may direct the dynamics of the system along a path other than that set by the exogenous initial conditions (though we may consider the structure of the system to be a kind of initial condition). The point is that the embodied plan gives freedom to the system, since it "might be otherwise'; there could be a different plan and a different outcome. We see this in the variety of life-forms: each follows its own algorithm of development, life-history and behaviour at the level of the individual organism. The existence of a plan as abstract pattern-information is a pre-requisite for options and therefore freedom of action.

2.5. Goal, Master Function and Will Nestedness

Now let us complete the connections between will, goal and function. The previous argument reveals an important difference between downward and any other kind of causation (considered important by Walker [58]): the former must always be directed by a purpose, for which we need to identify a goal (upward and same level causation are satisfactorily explained by initial conditions [48]). Viewing entities and actions both as the consequence of information constraining (filtering) entropic systems, then the role that is taken by initial conditions in upward causation, is taken by system-level pattern-information (the transcendent complex [50]) in downward causation. Since the goal G is a fixed point in an objective function $\mathcal{G}$, it constitutes information (e.g., a homoeostatic set-point) that must be embodied in the agent's internal organisation. Since the objective function $\mathcal{G}$ represents the overall functioning of the system (at its highest level), it matches the definition given by Cummins [43] and Farnsworth et al. [44]. The highest level function from which we identify the purpose of a system was termed the "master function" by Jaeger and Calkins [42], so the will of an agent is instantiated in the master function. This then identifies $\mathcal{G}$ with "master function" and "will" with $\mathrm{G}(\mathcal{G})$.

Ellis [48] identifies five types of downward causation, the second being "non-adaptive information control", where he says "higher level entities influence lower level entities so as to attain specific fixed goals through the existence of feedback control loops..." in which "the outcome is not determined by the boundary or initial conditions; rather it is determined by the goals". Butterfield [46] gives a more mathematically precise account of this, but without elaborating on the meaning or origin of "goals". Indeed, as both Ellis [48] and Butterfield [46] proceed with the third type: downward causation "via adaptive selection" they refer to fitness criteria as "meta-goals" and it is clear that these originate before and beyond the existence of the agent in question. Ellis [48] describes meta-goal as "the higher level "purpose" that guides the dynamics" and explains that "the goals are established through the process of natural selection and genetically embodied, in the case of biological systems, or are embodied via the engineering design and subsequent user choice, in the case of manufactured systems".

This suggests a nested hierarchy of goal-driven systems and for each, the goal is the source of causal power and as such may be identified as the "will" (free or otherwise). We may interpret the definition of "deep freedom" [13] as meaning that an agent has at least two nested levels of causal power, the higher of which, at least, is embodied within the agent (as causal pattern-information). This

concept may be formalised after introducing the discrete variable "will-nestedness" $\mathcal{N}$ which counts the number of levels of causal power exercised over a system, from within the agent as a whole (i.e., at the level of master function), the $\mathcal{N}$-th level being the highest-level internal cause of its actions.

Among organisms in general, the master function specifies the criteria by which the organism is to assess its possible future reactions to the environment. It is so much an integral part of the organism that without it, the organism would not exist. However, it was not chosen by the organism (in the sense of deep free-will) because it was created by evolutionary filtering and inherited from its parent(s); as all known life has been created by the previous generation copying itself. For single celled organisms the biological master function is to maximise their cell count by survival and reproduction, but in multicellular organisms, this master function exists, by definition, at the level of the whole organism (the unconstrained drive to proliferate a single cell line leads to cancer). Organisms with a central nervous system, regulated by neuro-hormone systems, with their corresponding emotions, can implement more complex (information rich) and adaptable (internally branched) algorithms for the master function, which may include will-nestedness $\mathcal{N} > 1$. In humans, this is taken to such an extent that the biological master function may seem to have been superseded (but the weight of socio-biological evidence may suggest otherwise [59,60]).

2.6. The Possibility of Choice and Alternative Futures

So far I have identified organisational-closure and the internal generation of a goal-based plan as prerequisites for free-will, but have not yet addressed the "alternative futures" problem relating to an agent constructed from elemental components that necessarily obey physical determinism. List [18] provides a philosophical argument for meeting this requirement, constructed from supervenience and multiple realisation of an agent in relation to its underlying (micro) physical level: "an agent-state is consistent with every sequence of events that is supported by at least one of its physical realizations" [18]. He shows that this may apply not only to multiple micro-histories up to t, but in principle includes subsequent $(t + \tau)$ sequences at the micro-level, which may map to different agent states and therefore permit different courses of action at the (macro) agent-level. To explain: for any given time t, the macro-state $Q_i(t)$ is consistent with a set of micro-states $\mathbf{s}$, at least one of which $s_i \in \mathbf{s}$ may lead (deterministically) to a new state $s_j \in \mathbf{s}$ at $t + 1$, with which a different macro-state Q_j is consistent, thus giving the agent a choice of which micro-state history to "ride" into $t + 1$ (this idea is developed with rigour by List [18], and illustrated with "real-world" examples; it is the basis on which he concludes that agents may be "free to do otherwise", despite supervening on deterministic physical processes).

Alternative futures may be produced at multiple levels of system organisation within a hierarchical structure, by re-applying the principles identified by List [18] for each level of macro-micro relations. For any system level L to have the potential for alternative futures, it must have the attributes of an "agent-level": supervenience and multiple realisation such that pattern-information with causal power emerges at level L from $L - 1$: i.e., a transcendent complex exists at level L. However, this does not necessarily give free will to a system of that level, since for that, it must be organisationally closed. If it were not so, we would not be able to identify the system at level L as an entity to which free-will could be ascribed. Thus will-nestedness cannot be attributed to levels of organisation below that of the Kantian whole. Since the Kantian whole is, by definition, the highest level of organisation to which free-will may be ascribed (any causal power beyond it rules out its free will), then will-nestedness can only apply at the level of the Kantian whole. Given this, the will-nestedness must be constructed from purely organisational, i.e., pattern-informational and therefore be purely computational in nature. This is an important deduction: free-will can only be an attribute of a Kantian whole and it can only result from the cybernetic structure at the level of the Kantian whole.

3. Choosing Possible Futures: The Computational Condition

We see that for free-will, an agent must have an independent and internally generated purpose for action and that this requires it to be organisationally closed. Free-will further requires the agent to use this purpose to choose among options. To do so, it needs an internal representation of possible futures from which to choose and an internally generated means of choosing. We now turn to the conditions which enable these essentially computational facilities.

3.1. Information Abstraction

The organisational boundary is where internal is distinguished from external. If the agent in question had no links of any kind between internal and external, then it would be unable to respond to, or use, its environment and in that case it could not be behaviourally autonomous as Froese et al. [22] defined it. Therefore to be both behaviourally and constitutively autonomous an agent must be disconnected from effective cause (also termed efficient cause), but able to perceive at least some aspect of its environment and act upon what it perceives (this being the foundation of cognition). The way this is achieved is via one or more transducer on the boundary, which acts as an intermediary between external and internal, allowing causal power without permitting effective cause to pass through the boundary. The transducer is a system which transforms information from one medium of embodiment to another (in analogy with an engine which transforms energy from one form to another). As information crosses the interface between one medium and another, it looses its effective causality. The reason is that information causality is mediated by physical forces. What is meant here by information causality, is that changes in an effective force (within a medium) constitute a pattern in force which as a pattern, constitutes information in the sense defined in Section 1. Forces are effective only within the medium to which they belong. The transducer has two sides, each belonging to a different medium, so each is causally linked to a different medium. Therefore fluctuations in force on one side exert no effective cause on the other side (because it is a different medium). However, the transducer allows a correlation between the fluctuation on one side and the other. What happens between the sides is a transformation of the information from one medium to the other and during that transformation the information looses its physical causal effect, but the transducer passes through its causal power via correlation, which itself may be modulated by the transducer. The disconnection of information from its medium is termed information abstraction because information without a medium of embodiment cannot really exist, it is an abstraction.

Information abstraction at the organisational boundary is crucial to achieving autonomy because it strips off the physical effect of the external environment to take only the abstract information which is then used as a signal in cognition. Causes are transformed into signals, their effects being rendered responses (which thereby may become optional). It is this separation of information (as signals) from the physical force of cause and effect that releases the agent from attachment to the cause-effect determinism of its environment. The material apparatus for performing this task is a transducer: the tegumental membranes of bacteria and other cells contain a wide variety of transducers (receptors) and we expect an artificial robot to be well equipped with them too. This is not a merely technical point. The closure condition gives the system a degree of causal independence and the boundary transducers give it a sensitivity to its environment whilst preserving this independence. The boundary is the place where the inevitability of cause and effect of the environment meets that of the internal processes of the system and the transducers are the interface between these causal chains. According to one interpretation, internally to the system, the environment is "reduced" to abstract, representational, information by the transducers [61]. Now the question is, what must the agent do with this information in order to exercise free-will?

Of course the answer is to compute: more specifically, to perform transformations on the data as a result of a sequence of physical changes in the physical structure of the agent. Such changes are described by automata theory, for which the most basic automaton has two states and can potentially change state on receiving a signal to which it responds: it is a switch (e.g., a protein molecule with two conformations is an "acceptor" of all strings from the alphabet $\{0, 1\}$, where these symbols may represent the presence/absence of e.g., another molecule or a level above a threshold (e.g., temperature)). Obviously, the switch is the elemental component for generating discrete options. A less obvious, but crucial property of the switch for the physical embodiment of computation is "thermodynamic indifference". Walker and Davies [62] focus on computation in explaining the origin of life, referring to genetics-first theories as "digital-first", emphasising the need for "programmability" and its provision by informational polymers (the genetic oligomers RNA, DNA etc.). By programmability they meant that components of a system are configured so that the system state can change reversibly, approximately independent of energy flow: i.e., changes of state are not accompanied by substantial changes in potential (stored) energy. If they were, then switching would always be biased by the difference in energetic cost between e.g., switching on and off. Energetically unbiased switching is the physical underlying mechanism of "information abstraction" referred to by Walker and Davies [62]. In reality, switching (and state-changes in general) always have energetic consequences (more deeply, there is always an exchange of entropy between the system and an external energy source), which is one of the reasons an autonomous agent must complete work cycles as Kauffman [63] specifies. What makes the informational polymers (e.g., DNA) of life special is the fact that they are reversible in a way that is thermodynamically indifferent (or very nearly so; see Ptashne [64]). Any ordered set of n switch positions (e.g., 1,1,0,0) has very nearly the same potential energy as any other ordered set of n (e.g., 0,1,0,1).

For free-will, both the self-assembly of autocatalytic systems and the computational requirements (switches and memory) are jointly necessary. Walker and Davies [62] and Walker [58] proposed that the autonomy of living systems arises from the combination of "analogue" chemical networks and "digital information processing". Accordingly, a hybrid automaton is a good model for the construction of a free-will agent. Figure 1 (from Hagiya et al. [10]) shows an example of a hybrid automaton which combines discrete-state with continuous dynamical systems, such that the discrete states (as modes of functioning) determine the system's responses to dynamic variables and these responses potentially influence the trajectory of the dynamics. In this pedagogical example from Hagiya et al. [10], there are two state variables α and τ which determine the set points for autonomous chemotaxis behaviour in the bacterium represented. This system can freely maximise its (experienced) environmental concentration of x (e.g., glucose), so its goal $\mathrm{G}(\mathcal{G})$ is defined, but it cannot choose how (hence $\mathcal{N} = 1$), so it cannot express free-will in the deep sense. However, if it had independent control over α and τ, with the ability to adjust these values according to a plan of its own making, then it would fulfil the condition of having willed its own behaviour ($\mathcal{N} = 2$), at least in the sense defined in the previous section. As part of a living bacterium, the system depicted would be a component of a Kantian whole (the free-living organism), so all that is missing for free-will is a plan for determining α and τ according to an internally generated goal (a master function) and some means of computing this. The computational requirements for free-will are identified as follows.

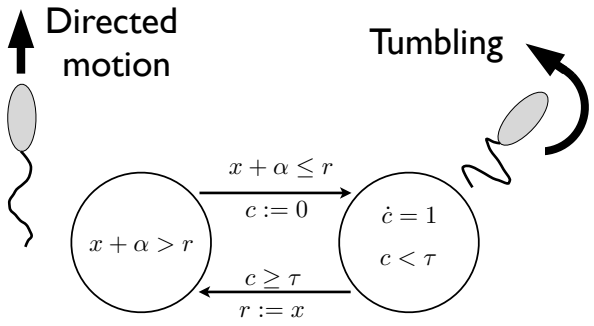

Figure 1. A bacterial chemotaxis controller described as a hybrid automaton, realised in practice by e.g., Escherichia coli species, but also as an engineering design for a molecular robot, using DNA-based components by Hagiya et al. [10], from which the figure, slightly modified, is taken. Note that α and τ are internal set-points, which constitute pattern-information embodied in the molecular robot's structure and are ultimately determined by Darwinian evolution (natural robots) or intentional design (human artefacts). The bacterium responds to the environmental concentration (e.g., of glucose); its objective function x, which is detected by a membrane transducer which generates the internal variable r. It is searching for a higher concentration when x is below a set threshold (directed motion) and tumbles randomly for a time set internally by τ using an internal "clock" signal c, whenever the concentration is at least equal to the threshold. The swimming and tumbling are modes of action of its flagella. The objective function of this system is the experienced concentration x and the goal is $\max(x)$. The chemotaxis controller exhibits will-nestedness of one.

3.2. The Representation of Self

Firstly, a free-will agent must maintain an internal representation of itself, and also the effect of its environment on its internal state, to enable it to assess each of its options for action.

At the root of automata theory lies an attempt to fully describe (and therefore predict) the behaviour of a system without a detailed mechanistic account of its internal workings (the black box approach). The system is captured in the mapping between environmental stimuli and responses:

$$R(t+1) = W(\mathbf{S}(t), S(t)) \tag{2}$$

where $R(t)$ and $S(t)$ are the response and stimulus, and $\mathbf{S}(t)$ is the history of stimuli experienced by the system, from the beginning of its formation up to t and $W(\cdot)$ is the mapping function. This presents an immediate problem, since in general $\mathbf{S}(t)$ is arbitrary and infinite in range (it is instructive to think of Strawson's [20] "Basic Argument" in terms of Equation (2): the response of a system, in general, depends on its environment from before the system came into existence). The solution to this indefinite $\mathbf{S}(t)$ problem (provided by Moore [65]) is to assume that the infinite set of $\mathbf{S}(t)$ may be partitioned into a finite number of disjoint equivalence sets, each containing the histories that are equivalent in their effect on $R(t+\tau)$, $\forall\tau$, where τ is in the interval $[0, \infty]$. These equivalence sets are represented as states $Q(t)$ of the system, so that:

$$R(t+1) = W(Q(t), S(t)) \text{ and } Q(t+1) = U(Q(t), S(t)), \tag{3}$$

where $U(Q(t), S(t))$ is the transformation function dictating the transition of system state given its present state and that of the stimulus (see Minsky [66], pp. 16–17). Thus $Q(t)$ represents how the system is now, given its previous history of experiences. $Q(t)$ corresponds to the agent-state of List [18], which is multiply-realisable and which is, at least in principle, free to take more than one value at some point in the future $t+\tau$, despite supervening on a a wholly determined set of micro-histories. List [18] showed the possibility of choice at the agent-level, but this does not necessarily mean that $Q(t)$ is indeterminate. Specifically, for free-choice (of $Q(t+1)$ and implied $R(t+1)$), the direction

taken at the branching point t must be determined by a process internal to the agent that represents its "purpose" (as defined earlier). For the choice to be purposeful, it must be based on an assessment of the outcomes that would arise from choosing each of the options. This entails a prediction of possible futures, for which a free-will system must have a model of itself in its environmental context.

The question now is, how can a system create such a representation by and for itself, not "programmed" by some exogenous source of information? The answer seems to be as it is with material self-assembly: a boot-strap, step by step, gathering of pattern-information embodied in form, such that as the form grows, it increases in complexity. In the particular case of building a model of self and environment, this process is one of learning, for which the field of "machine learning" provides our understanding. Well known advances in this field have already led to sophisticated learning among pre-existing (i.e., not self-assembled) computation systems such as deep neural networks etc. The difference here is that the learning is not merely a statistical problem, but one of simultaneous self-construction, which must begin with simple systems, so in the remainder of this section, only basic and simple systems capable of unsupervised learning are discussed.

3.2.1. Learning in a Constant Environment

The most basic form of learning is operant conditioning (reinforcement learning), described mathematically by Zhang [67] (cited in Krakauer [68]) as follows from the description provided by [68]. Let R_i be one of a set of N possible responses ($R_i \in \mathbf{R}, i \in [1, N]$) in a constant environment, occurring with probability r_i. For each response there is a "reward" ρ (which coincides with the objective function $\mathcal{G}$ that defines the goal of the system: $\rho \rightarrow f(\mathcal{G})$), so that the incremental change in probability of the i-th response is:

$$\Delta r_i = a\rho_k(\kappa_{k,i} - r_i), \tag{4}$$

where $\kappa_{k,i}$ is the Kronecker delta function (equal to 1 if $i = k$, else equal to zero) and a is a learning rate constant. Since the average change in response over the ensemble of possible responses is the frequency-weighted sum: $\Delta\widehat{r} = \sum_k^N r_k \Delta r_k$, the result is that the frequency of the i-th response incrementally increases in proportion to the difference between its reward and the average over all rewards:

$$\Delta r_i = a r_i(\rho_i - \widehat{\rho}) \text{ , where } \widehat{\rho} = \sum_k^N r_k \rho_k. \tag{5}$$

The dynamic quantified in Equation (5) describes learning by maximising the reward experienced. Such learning is equivalent to making an increasingly accurate model of the (static) relationship between the agent's internal state and the environment, via (Bayesian) "trial and error" sampling of responses. Given a constant environment, the solution to Equation (5) yields a single, reward maximising response: $R^* = R_i$ such that $\rho_i = \widehat{\rho}$ and $r_i = \kappa_{k,i}$.

To achieve this in practice the system must at least keep a record of the reward for the last response made and the average reward, for which an automaton with an external memory is required (e.g., a push-down automaton, though this is still essentially a determinate finite state automaton DFA). Quantifying the complexity of such a system is probably best achieved through a programme (algorithmic) complexity measure since the information instantiated by such an automaton is almost all in its transition mapping and there are robust methods for reducing this to the minimum description, leading directly to the Kolmogorov complexity. The process of learning can be interpreted in information terms: if the starting probability distribution of $\mathbf{R}$ is $\tilde{r}$, then the initial Shannon entropy of the system is $H = -\sum_i^N \tilde{r}_i \log(\tilde{r}_i)$ and the final entropy is zero: having completed its learning, the system has no uncertainty about the best way to respond to this environment. In this state, the automaton is a complete representation of its interaction with its environment (i.e., the distribution

of rewards over its repertoire of responses) and it embodies exactly H units of information: a quantity which should match the algorithmic complexity measure (though not tested here).

3.2.2. Extension to a Variable Environment

Generalising to a variable environment, for which a set of finite states $\mathbf{S}$ is an adequate representation, there would exist a reward maximising response for each state: $R_i^* \rightarrow (S_i)$, such that R_i^* solves Equation (5). The agent may choose to maximise its reward over all $\mathbf{S}$, and to enable this, it must learn the best response for every $S_i \in \mathbf{S}$. In information terms, first let $H(\mathbf{s})$ be the entropy of the environment having probability distribution $\mathbf{s}$ and $H_t(\mathbf{r_t})$ be that of the responses, given their probability distribution $\mathbf{r_t}$ at time t. The Shannon information the agent has about its environment (in terms of its rewards) is:

$$\begin{aligned} I_t(\mathbf{s} : \mathbf{r_t}) &= H(\mathbf{s}) + H(\mathbf{r_t}) - H(\mathbf{s}, \mathbf{r_t}) \\ &= H(\mathbf{s}) + H(\mathbf{r_t}) - H(\mathbf{s}|\mathbf{r_t}), \end{aligned} \tag{6}$$

meaning that the agent is learning both the distribution $\mathbf{s}$ and the reward associated with each $S_i \in \mathbf{S}$. This mutual information is embodied in the structure of the agent and can be used as a measure of its complexity. The structure of the agent may be too simple to embody as much as the maximum mutual information, in which case its learning will be limited and it will not make optimal responses, so Equation (6) is a measure of the minimum complexity required for optimal behaviour from the agent. It would be possible for an agent to implement this learning system by "growing" multiple copies of the DFA with memory (one for each $S_i \in \mathbf{S}$) that is used for a constant environment. The output of each of these would then be the input to a further DFA which it uses to maximise the reward across all of them. This "growth" would be enacting meta-learning: the agent would increase its complexity in response to rewards. The number of states in $\mathbf{S}$ is not known by the agent a-priori, so the number of DFAs needing to be "grown" is indeterminate. Further, account should be taken of the extended time needed to perform such laborious learning and the consequences of the agent being wrong in so many trials. It seems that for practical reasons there comes a point when a more powerful kind of computation becomes necessary. In computation terms, such a learning problem requires at least a finite and non-volatile memory, effectively to store multiple instances of the single learning problem encountered in a constant environment. For this reason a Turing Machine would be a more realistic option.

3.3. The Free-Will Machine

These computational requirements for free-will to become possible are brought together in the hypothetical "free-will machine" of Figure 2. This information processing must be implemented by the agent to which free-will is ascribed and that agent must, further, be a Kantian whole for the requirements of free-will to met. The current state of the environment (external) and of the agent (internal) are derived by information abstraction from the physical world: the array of receptor molecules in the cell membrane, the nervous senses of an animal, or the transducers of a human artefact all perform this task. The first Turing machine implementation TM1 constitutes a representation of the agent in the present (relevant aspects of its environment are included) and is informed (updated) by the state information Q_t and S_t. The function of this representation is to identify the set of possible responses $\underline{R}_t$ that the agent can make, given Q_t, S_t (underline notation now denotes a set).

TM2 uses these hypothetical responses to compute the set of possible futures at a time $t + n$ (n may take any positive value) $\underline{F}_{t+n}$ for which in a simple case TM2 : $\underline{R}_t \rightarrow \underline{F}_{t+n}$ is a map of responses onto possible futures (simple, because there is not necessarily a $1:1$ relation between $\underline{R}_t$ and $\underline{F}_{t+n}$, but we need not be concerned with that at present). There is no limit on the number of possible futures that TM2 may compute, but it must be at least 2 for a choice to exist. These possible futures are each represented by a set of states $\underline{f}$, each member f_n being equivalent to a prediction of a possible Q_{t+n}.

The finite state automaton FSA chooses from among these, using a selection criterion based on the objective function defined by a goal G, which is generated by the agent (not exogenously). This goal is the maximisation of the master function, (e.g., for a living agent this is life-time reproductive success). The goal enables the optimal possible future state to be recognised (it is the one which maximises the master function) and this future state f' implies an optimal response r' (in general there could be more than one, in which case the agent will be indifferent among them). Having selected an optimal response, the agent then must implement it in the physical realm. Since the computation of r' has been conducted in the realm of information, this step appears to involve the control of material by information. In practice all the computation and indeed all the information is instantiated by material and energy acting in the physical realm, so our cybernetic model is merely an abstract representation of the organisation of the physical processes which lead to the implementation in the physical system and this return to physical reality is represent by the action (IPS). Such implementation inevitably results in a transformation of the agent into a new state Q_{t+1}, together with S_{t+1} and this restarts the cycle. It may be noted that Von Neumann's self-replicating automata are proven universal Turing machines [39] and Turing machines are thought to be common among living systems, so this computational arrangement is not beyond the bounds of possibility.

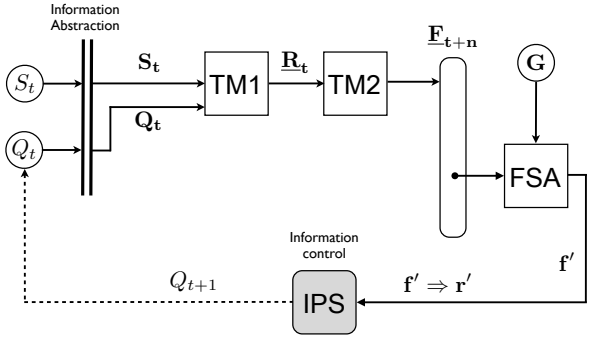

Figure 2. A conceptual "free-will" machine, which generates predictions of its state in alternative futures $\underline{F}_{t+n}$using an internal representation of itself interacting with its environment, selecting the optimal from among these, using a goal-bases criterion G, in which the goal is internally determined (further explanation in the text).

4. Discussion and Synthesis

To summarise, the essential requirements for free will are:

R1 There must be a self to self-determine.
R2 There must be a non-zero probability of more than one option being enacted.
R3 There must be an internal means of choosing among options (which is not merely random, since randomness is not a choice).

For R3 to be fulfilled:

R4 Options must be generated from an internal model of the self which can calculate future states contingent on possible responses.
R5 Choosing among these options requires their evaluation using an internally generated goal.
R6 For "deep free-will", at least two nested levels of choice and goal (R4–R5) must be enacted by the agent.

R1 and references to "internally generated" are fulfilled by organisational closure. For R2, the possibility of options, which implies "multiple futures" has been established for the level of the agent by List [18]. R3 and its predicates R4–R5 imply a minimum level of computational power, which in principle can be met by a small set of Turing machines, which may in principle be implemented

by a von Neumann architecture computer, a network or cellular automaton-based or any other sort of computer, including a biomolecular system such as found in higher animal life, but it seems to be beyond the power of a single living cell (though that last point is not yet established). R4 in particular seems to require a finite memory (the size depends on the complexity of the agent and its environment) and R6, the qualifier for deep free-will, adds a little more to the computing power necessary, but it is important to note that this extra is not a step-change: it is not qualitatively more demanding than the automated decision making required by R3.

The question of free-will is not one of whether an agent's actions are caused, since all actions ultimately have a cause. The ultimate cause of any action can be understood as resulting from selection over random actions by a pattern, which leaves a correlation with the pattern that caused the selection (instantiating pattern-information). All living organisms, including people, were produced by information-pattern filtering, proximally by molecular replication (creating inheritance) and ultimately by Darwinian selection. All human artefacts were created by following a design pattern (though it may not have been completed before artefact construction), so they correlate with their design. Even inanimate objects, such as stars, lakes and sand grains, owe their form to the information-pattern of underlying physical laws, Pauli's exclusion principle and the distribution of matter and energy in space following the big-bang. To this, we must add randomness which has been entering as "informational raw material" into the universe, disrupting the original patterns and opening opportunities for novelty (evolutionary for life) and more widely directing the course of the universe in unexpected ways as its history tracks a course in the highly ergodic space of possibilities.

Taking Strawson's [20] Basic Argument seriously, this pattern-correlation and the injection of randomness both deny free will. From them, we obtain a model in which the identity of all things, including human beings, is an illusion: as if the universe was all one complex manifestation which only appears to include separate agents. Closer inspection shows how the nested-hierarchical construction of this complexity entails the creation of genuinely new pattern-information, caused by and embodied in the interactions among component parts of putative agents. This novel pattern-information transcends its component parts and can exert downward causation upon them. Some structures (such as autocatalytic sets) created this way are organisationally closed (though materially and thermodynamically remain open systems). Because of this, their internal dynamics are, at least partially, separated from the external dynamics of their environment and this gives them an organisational boundary, enabling internal to be defined against external. At this boundary, external and internal chains of cause and effect interact through transducers which transform physical determinism into stimulus-response relations. Systems with these properties are essentially cybernetic and although their low-level processes are continuous with the rest of the universe, List [18] has shown that in principle they may have options for their next state and response: they are freed from physical determinism. To translate this freedom into free-will, requires that the (partially) independent agent chooses from among its options and this entails an internal computation of possible futures and their evaluation against a goal representing the fulfilment of the agent's "master function" (i.e., its purpose). This goal is a fixed point in an objective function which may be simple (as in a homoeostatic system), but also arbitrarily complex and multi-layered, taking account of multiple time-scales and interactions with other agents. If the objective function is at least two-layered (will-nestedness $\mathcal{N} \geq 2$), then it effectively has a choice of what to choose and thereby could fulfil the established definition of (deep) free-will [13,16]. This calls into question the idea that free-will is an all or nothing capacity, instead, it suggests free-will to be a discrete quantity and even something we could in principle measure as a trait of a system.

The reason is that deep free-will has so far been defined as the freedom to choose ones will, but the analysis presented here shows that to be wilful, a choice must be purposeful, which means optimising an objective function. Freedom is in the choice of objective function. Since therefore, the core of will is the objective function, deep freedom is the freedom to choose this, but to be wilful, this choice in turn must optimise a hierarchically superior objective function, which must have been

determined by something. We can conceive of a large but finite nested set of such objective functions, but ultimately the highest of them all must be provided either arbitrarily (e.g., at random) or by natural selection (or its unnatural equivalent), or by design: in all cases, not the free choice of the agent. This applies even to human beings, who are still subject to "design" by inheritance, selected by evolution. The depth of free-will is therefore a discrete, finite variable (will-nestedness $\mathcal{N}$) and we may speak of one kind of agent being more deeply free than another, but no kind of agent can be ultimately free-willed in the sense meant by deep free-will (presumably, humans attain the highest will-nestedness of known agents).

Seeking an ultimate cause of will leads to an infinite regress because for (second order) will-setting to be "rational" (meaning a reason underlies it), it must be based on the setting of a (higher order) will. This introduces a third order rational basis of the will-setting and so on ad infinitum. Here the problem is explained in the more concrete and formal terms of fixed points (goals) in objective functions. The concept is made sufficiently specific to quantify (as nestedness) and used to conclude that no agent can be ultimately free willed in the strong source-theory sense of being ultimately responsible. Philosophers (such as Strawson) conclude that if the ultimate source is not of the agent, then the agent cannot have free will. The more quantitatively oriented analysis presented here adds a nuance to this, so we are not forced into an "all or nothing" conclusion about free will. It suggests that agents may be better characterised by the degree of free-will, expressed in terms of the will-nestedness, which can in principle be derived for any kind of agent under scrutiny. Therefore, whilst Frankfurt [12] asserts that only humans can have second order volition, but animals can have first order, I contend that any autonomous agent (including synthetic) can be characterised by the hierarchical order of their goal setting (be it zero, one, two or higher) and further that this order can be identified as the number of hierarchically arranged objective functions in their decision making (computation) that are embodied (as information) within their structure (what I call will nestedness).

If interpreting the philosopher's definition of deep free-will as $\mathcal{N} \geq 2$ is correct, then it is not hard to achieve in principle. The impediment to a non-life robot acquiring that sort of free-will is not computational, it is the closure constraint with its requirement for bootstrapping self-assembly (especially since this includes the "growth" of the sort of computational apparatus indicated in Figure 2). That is a problem already solved by life. Quite likely $\mathcal{N} \geq 2$ in most or all organisms having at least a limbic system, so free-will defined this way can be attributed to most or all vertebrates [69]. The computational requirements for exercising purposeful choices are not very challenging for artificial computers. Among human artefacts, including what most people would define as robots, the organisational closure condition is the major hurdle not yet leapt. This point has been recognised in the artificial intelligence literature, especially concerning "cognitive robotics" which emphasises embodiment [22,26]. For the time-being, it seems free-will, as defined here, is a unique property of living things, but the possibility of extending it to synthetic robots remains.

Acknowledgments: This work was unfunded, but was inspired by attendance at the workshop "Information, Causality and the Origin of Life" in Arizona State University at Tempe, AZ 30 September–2 October 2014, funded by the Templeton World Charity Foundation. The cost of open access publication was met by the Queen's University Belfast.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DFA	Determinate Finite State Automaton
DM	Determinate Machine
FSA	Finite State Automaton
NDM	Non-Determinate Machine
TM	Turing Machine
UTM	Universal Turing Machine

References

1. Sharov, A.A. Functional Information: Towards Synthesis of Biosemiotics and Cybernetics. *Entropy* **2010**, *12*, 1050–1070.
2. Hazen, R.M. The emergence of patterning in life's origin and evolution. *Int. J. Dev. Biol.* **2009**, *53*, 683–692.
3. Farnsworth, K.; Nelson, J.; Gershenson, C. Living is Information Processing: From Molecules to Global Systems. *Acta Biotheor.* **2013**, *61*, 203–222.
4. Von Weizsäcker, C.F. *Die Einheit der Natur*; Deutscher Taschenbuch Verlag: Munich, Germany, 1974.
5. Stonier, T. Information as a basic property of the universe. *Biosystems* **1996**, *38*, 135–140.
6. Devlin, K.J. *Logic and Information*; Cambridge University Press: Cambridge, UK, 1992.
7. Floridi, L. Information. In *The Blackwell Guide to the Philosophy of Computing and Information*; Floridi, L., Ed.; Blackwell Publishing Ltd.: Hoboken, NJ, USA, 2003; pp. 40–61.
8. Szostak, J.W. Functional information: Molecular messages. *Nature* **2003**, *423*, 689.
9. Hazen, R.M.; Griffin, P.L.; Carothers, J.M.; Szostak, J.W. Functional information and the emergence of biocomplexity. *Proc. Natl. Acad. Sci.* **2007**, *104*, 8574–8581.
10. Hagiya, M.; Aubert-Kato, N.; Wang, S.; Kobayashi, S. Molecular computers for molecular robots as hybrid systems. *Theor. Comp. Sci.* **2016**, *632*, 4–20.
11. Pfeifer, R.; Bongard, J. *How the Body Shapes the Way We Think: A New View of Intelligence*; MIT Press: Boston, MA, USA, 2007.
12. Frankfurt, H. Freedom of the will and the concept of a person. *J. Philos.* **1971**, *68*, 5–20.
13. Kane, R. *A Contemporary Introduction to Free Will*; Oxford University Press: Oxford, UK, 2005.
14. McKenna, M.; Pereboom, D. *Free Will: A Contemporary Introduction*; Routledge: Abingdon, UK, 2016.
15. Coeckelbergh, M. *The Metaphysics of Autonomy*; Palgrave Macmillan: London, UK, 2004.
16. Westen, P. Getting the Fly out of the Bottle: The False Problem of Free Will and Determinism. *Buffalo Crim. Law Rev.* **2005**, *8*, 599–652.
17. Van Inwagen, P. *An Essay on Free Will*; Oxford University Press: Oxford, UK, 1983.
18. List, C. Free will, determinism, and the possibility of doing otherwise. *Noûs* **2014**, *48*, 156–178.
19. Van Inwagen, P. Some Thoughts on An Essay on Free Will. *Harvard Rev. Phil.* **2015**, *22*, 16–30.
20. Strawson, G. *Freedom and Belief*; Oxford University Press: Oxford, UK, 1986.
21. Kauffman, S.A. Autocatalytic sets of proteins. *J. Theor. Biol.* **1986**, *119*, 1–24.
22. Froese, T.; Virgo, N.; Izquierdo, E. Autonomy: A review and a reappraisal. In Proceedings of the European Conference on Artificial Life, Lisbon, Portugal, 10–14 September 2007; pp. 455–464.
23. Varela, F. *Principles of Biological Autonomy*; Elsevier: Amsterdam, The Netherlands, 1979.
24. Zeleny, M. What is autopoiesis? In *Autopoiesis: A Theory of Living Organization*; Elsevier North Holland: New York, NY, USA, 1981; pp. 4–17.
25. Rosen, R. *Life Itself*; Columbia University Press: New York, NY, USA, 1991.
26. Vernon, D.; Lowe, R.; Thill, S.; Ziemke, T. Embodied cognition and circular causality: On the role of constitutive autonomy in the reciprocal coupling of perception and action. *Front. Psychol.* **2015**, *6*, doi:10.3389/fpsyg.2015.01660.
27. Bich, L. Systems and organizations: Theoretical tools, conceptual distinctions and epistemological implications. In *Towards a Post-Bertalanffy Systemics*; Springer International Publishing: Cham, Switzerland, 2016; pp. 203–209.
28. Varzi, A. Mereology. Available online: http://philsci-archive.pitt.edu/12040/ (accessed on 19 May 2017).
29. Effingham, N. *An Introduction to Ontology*; Polity Press: Cambridge, UK, 2013.
30. Van Inwagen, P. When Are Objects Parts? *Phil. Perspect.* **1987**, *1*, 21–47.
31. Heylighen, F. Relational Closure: A mathematical concept for distinction-making and complexity analysis. *Cybern. Syst.* **1990**, *90*, 335–342.
32. Kauffman, S.; Clayton, P. On emergence, agency, and organization. *Biol. Philos.* **2006**, *21*, 501–521.
33. Ginsborg, H. Kant's biological teleology and its philosophical significance. In *A Companion to Kant*; Bird, G., Ed.; Blackwell: Oxford, UK, 2006; pp. 455–469.
34. Hordijk, W.; Steel, M. Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *J. Theor. Biol.* **2004**, *227*, 451–461.
35. Hordijk, W.; Hein, J.; Steel, M. Autocatalytic Sets and the Origin of Life. *Entropy* **2010**, *12*, 1733–1742.

36. Segré, D.; Ben-Eli, D.; Lancet, D. Compositional genomes: Prebiotic information transfer in mutually catalytic noncovalent assemblies. *Proc. Natl. Acad. Sci.* **2000**, *97*, 4112–4117.
37. Luisi, P. Autopoiesis: A review and a reappraisal. *Naturwissenschaften* **2003**, *90*, 49–59.
38. Davies, J.A. *Life Unfolding: How the Human Body Creates Itself*; Oxford University Press: Oxford, UK, 2014.
39. Von Neumann, J.; Burks, A. *Theory of Self-Reproducing automata*; Illinois University Press: Chicago, IL, USA, 1966.
40. Varela, F.; Maturana, H.; Uribe, R. Autopoiesis: The organization of living systems, its characterization and a model. *Curr. Mod. Biol.* **1974**, *5*, 187–196.
41. Auletta, G.; Ellis, G.F.R.; Jaeger, L. Top-down causation by information control: From a philosophical problem to a scientific research programme. *J. R. Soc. Interface* **2008**, *5*, 1159–1172.
42. Jaeger, L.; Calkins, E.R. Downward causation by information control in micro-organisms. *Interface Focus* **2012**, *2*, 26–41.
43. Cummins, R. Functional Analysis. *J. Philos.* **1975**, *72*, 741–765.
44. Farnsworth, K.D.; Albantakis, L.; Caruso, T. Unifying concepts of biological function from molecules to ecosystems. *Oikos* **2017**, doi: 10.1111/oik.04171.
45. Lorenz, D.; Jeng, A.; Deem, M. The emergence of modularity in biological systems. *Phys. Life Rev.* **2011**, *8*, 129–160.
46. Butterfield, J. Laws, causation and dynamics at different levels. *Interface Focus* **2012**, *2*, 101–114.
47. List, C. Levels: Descriptive, Explanatory, and Ontological. Available online: http://philsci-archive.pitt.edu/12040/ (accessed on 19 May 2017).
48. Ellis, G.F.R. Top-down causation and emergence: Some comments on mechanisms. *Interface Focus* **2012**, *2*, 126–140.
49. Ellis, G.F.R. On the nature of causation in complex systems. *Trans. R. Soc. S. Afr.* **2008**, *63*, 1–16, doi:10.1080/00359190809519211.
50. Farnsworth, K.D.; Ellis, G.F.R.; Jaeger, L. Living through Downward Causation. In *From Matter to Life: Information and Causality*; Walker, S.I., Davies, P.C.W., Ellis, G.F.R., Eds.; Cambridge University Press: Cambridge, UK, 2017; pp. 303–333.
51. Gilbert, S. *Developmental Biology*; Sinauer Associates: Sunderland, MA, USA, 2013.
52. Seeley, T. *Honeybee Democracy*; Princeton University Press: Princeton, NJ, USA, 2010.
53. Lineweaver, C.H.; Egan, C. Life, gravity and the second law of thermodynamics. *Phys. Life Rev.* **2008**, *5*, 225–242.
54. Adami, C.; Ofria, C.; Collier, T. Evolution of biological complexity. *Proc. Natl. Acad. Sci.* **2000**, *97*, 4463–4468.
55. Hoel, E.; Albantakis, L.; Marshall, W. Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neurosci. Conscious.* **2016**, *1*, niw012.
56. Marshall, W.; Albantakis, L.; Tononi, G. Black-boxing and cause-effect power. *arXivs* **2016**, arXiv:1608.03461.
57. Mayr, E. Teleological and Teleonomic: A New Analysis. In *Methodological and Historical Essays in the Natural and Social Sciences*; Springer: Dordrecht, The Netherlands, 1974; pp. 91–117.
58. Walker, S.I. Top-down causation and the rise of information in the emergence of life. *Information* **2014**, *5*, 424–439.
59. Wilson, D.; Wilson, E.O. Rethinking the Theoretical Foundation of Sociobiology. *Q. Rev. Biol.* **2007**, *82*, 327–348.
60. Wilson, E.O. *Sociobiology: The New Synthesis*; Harvard University Press: Cambridge, MA, USA, 1975.
61. Danchin, A. Bacteria as computers making computers. *FEMS Microbiol. Rev.* **2009**, *33*, 3–26.
62. Walker, S.; Davies, P. The algorithmic origins of life. *J. R. Soc. Interface* **2013**, *10*, 20120869.
63. Kauffman, S.A. *Investigations*; Oxford University Press: Oxford, UK, 2000.
64. Ptashne, M. Principles of a switch. *Nat. Chem. Biol.* **2011**, *7*, 484–487.
65. Moore, E.F. Gedanken-experiments on sequential machines. *Auto. Stud.* **1956**, *34*, 129–153.
66. Minsky, M.L. *Computation–Finite and Infinite Machines*; Prentice Hall: Englewood Cliffs, NJ, USA, 1967.
67. Zhang, J. Adaptive learning via selectionism and Bayesianism, Part 1: Connection between the two. *Neural Netw.* **2009**, *22*, 220228.

68. Krakauer, D. The inferential evolution of biological complexity: Forgetting nature by learning nurture. In *Complexity and the Arrow of Time*; Lineweaver, C.H., Davies, P.C.W., Ruse, M., Eds.; Cambridge University Press: Cambridge, UK, 2013; pp. 224–245.
69. Bruce, L.L.; Neary, T.J. The Limbic System of Tetrapods: A Comparative Analysis of Cortical and Amygdalar Populations. *Brain Behav. Evol.* **1995**, *46*, 224–234.